Selfinformative Limits of Bayes Estimates and Generalized Maximum Likelihood

Olaf Bunke and Jan Johannes
Humboldt University, Berlin

Summary: A definition of selfinformative Bayes carriers or limits is given as a description of an approach to noninformative Bayes estimation in non- and semiparametric models. It takes the posterior w.r.t. a prior as a new prior and repeats this procedure again and again. A main objective of the paper is to clarify the relation between selfinformative carriers or limits and maximum likelihood estimates (MLE's).
For a model with dominated probability distributions we state sufficient conditions under which the set of MLE's is a selfinformative carrier or in the case of a unique MLE its selfinformative limit property. Mixture models are covered. The result on carriers is extended to more general models without dominating measure.
Selfinformative limits in the case of estimation of hazard functions based in censored observations and in the case of normal linear models with possibly nonidentifiable parameters are shown to be identical to the generalized MLE's in the sense of Gill (1989) and Kiefer and Wolfowitz (1956). Selfinformative limits are given for semiparametric linear models. For a location model they are identical to generalized MLE's, while this is not true in general.

AMS 1980 subject classifications. Primary 62F G05; Secondary 62A15.

Key words and phrases. generalized maximum likelihood, Dirichlet prior, mixture models, semiparametric models, multivariate linear models, hazard functions, censoring.

* The research on this paper was carried out within the Sonderforschungsbereich 373 "'Quantifikation und Simulation ökonomischer Prozesse"' at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

1 Introduction

Bayes estimates are not only useful when some prior information is incorporated into the statistical analysis, but also as a basis for the construction of sensible estimates without prior information by a data dependent adaptation of parameters of the prior (probability) distribution (see Lindley and Smith (1972), Efron and Morris (1973)). Moreover there are attempts to choose a "noninformative" prior and (or) corresponding posterior distribution in view of a situation without prior information (see Hartigan (1983), Box and Tiao (1992)).
An alternative approach to the description of a noninformative situation was introduced in Bunke (2002). The idea is to give in some sense an infinitely increasing weight to the information contained in the observations in comparison to the prior distribution. This is done by taking the posterior as a new prior and calculating (for the same observations) the corresponding posterior and repeating the procedure again and again. If the Bayes estimate calculated in this manner converges to a limit, we will call it a selfinformative limit. Such a limit will be a sensible estimate in a noninformative situation. Indeed, when the observations follow a model with densities depending in finite dimensional parameters the selfinformative limit exists and is the maximum likelihood estimate (MLE). This was shown in Bunke, Hennig, and Schmidt (1976) under some regularity conditions ensuring the uniqueness of the MLE.
Our paper is devoted to the characterization of selfinformative limits under more general assumptions, e.g. when MLE are not unique or when they even don't exist as e.g. in semi- or nonparametric models.
In section 2 we will define selfinformative carriers and limits in a precise manner. We treat in section 3 a model with densities, which includes the parametric case and a semi- or nonparametric mixture case. The MLE and selfinformative limits appear to be identical. The sections 4 and 5 are directed to further cases in which the selfinformative estimates and the generalized MLE are identical: a normal linear regression model and the estimation of a hazard function. Section 6 is devoted to the determination of selfinformative limits in a general semiparametric multivariate linear model with normal-gamma priors for the unknown parameters and Dirichlet distribution for the unknown error distributions generalizing the results in Bunke (2002). As seen in section 7 these selfinformative limits turn out to be in general different from generalized MLE, which we calculate following the definitions of Kiefer and Wolfowitz (1956). The structure of the selfinformative limits has a sound intuitive background and their form is simple and resembles standard estimates in difference to the somewhat degenerated form of the generalized MLE. Therefore these "selfinformative" estimates deserve special attention and further investigations especially also replacing Dirichlet distributions by other.

2 Selfinformative limits

We consider a Bayes approach to the estimation of the parameter θ in a model for the random variable $\boldsymbol{X}$:

$$\boldsymbol{X} \sim P_\theta \quad , \quad \theta \in \Theta. \tag{2.1}$$

where P_θ is a probability distributions (p.d.) depending on an unknown parameter θ.

For this we assume $(\boldsymbol{\theta}, \boldsymbol{X})$ to be a random variable with values in $\Theta \times \mathcal{X}$ and to have a probability distribution $P^{\boldsymbol{\theta},\boldsymbol{X}}$ on a product σ-Algebra $\mathcal{B} \times \mathcal{A}$. For simplicity we assume Θ and $\mathcal{X}$ to be complete separable metric (polish) spaces and $\mathcal{B}, \mathcal{A}$ to be the corresponding σ-algebras of Borel sets.
The regular conditional distribution (c.p.d.) $P^{\boldsymbol{X}\,|\,\boldsymbol{\theta}=\theta} = P_\theta$ of $\boldsymbol{X}$ under the condition $\boldsymbol{\theta} = \theta$ will be the p.d. in the model (2.1). The marginal p.d. $\xi = P^{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is called the prior p.d., while the c.p.d. $\xi_x = P^{\boldsymbol{\theta}\,|\,\boldsymbol{X}=x}$ is called the posterior p.d. . The posterior ξ_x may be used in different ways for inferences on $\boldsymbol{\theta}$ based on the observation x of $\boldsymbol{X}$ (see Hartigan (1983) or Bernardo and Smith (1994)). If it possible to define a "posterior mean"

$$\hat{\theta}(x) := \int \theta \; \xi_x(d\theta) \quad , \tag{2.2}$$

it may be used as a sensible estimate of the unknown θ.
Now we will give a precise description of the stepwise approach, which uses iteratively the posterior as a new prior in the same model and repeats this procedure using always the given observation x:

1. step: Take random variables $\boldsymbol{\theta}_1, \boldsymbol{X}_1$ with the marginal p.d. ξ of $\boldsymbol{\theta}_1$ and the c.p.d P_θ of $\boldsymbol{X}_1$ under $\boldsymbol{\theta}_1 = \theta$. Choose a determination of the c.p.d of $\boldsymbol{\theta}_1$ under the condition $\boldsymbol{X}_1$ and take the special value $\boldsymbol{X}_1 = x$:

$$\xi_x^{(1)} = P^{\boldsymbol{\theta}_1\,|\,\boldsymbol{X}_1=x}. \tag{2.3}$$

2. step: Take random variables $\boldsymbol{\theta}_2, \boldsymbol{X}_2$ with the marginal p.d. $\xi_x^{(1)}$ of $\boldsymbol{\theta}_2$ and the c.p.d. P_θ of $\boldsymbol{X}_2$ under $\boldsymbol{\theta}_2 = \theta$. Choose a determination of the c.p.d. of $\boldsymbol{\theta}_2$ under the condition $\boldsymbol{X}_2$ and take

$$\xi_x^{(2)} = P^{\boldsymbol{\theta}_2\,|\,\boldsymbol{X}_2=x}. \tag{2.4}$$

Repeating this we arrive after r-steps at our r-th iteratively determined posterior $\xi_x^{(r)}$. The r-th posterior $\xi_x^{(r)}$ will depend on the chosen determinations of the c.p.d.'s

$P^{\boldsymbol{\theta}_j \,|\, \boldsymbol{X}_j = x}$. Assume now random variables $\boldsymbol{\theta}^{(r)}, \boldsymbol{X}^{(r)} = (\boldsymbol{X}_1^{(r)}, \ldots, \boldsymbol{X}_r^{(r)})$, where the marginal p.d. of $\boldsymbol{\theta}^{(r)}$ is the prior ξ and conditionally on $\boldsymbol{\theta}^{(r)} = \theta$ the random variables $\boldsymbol{X}_1^{(r)}, \ldots, \boldsymbol{X}_r^{(r)}$ are i.i.d. with distribution P_θ. If we would implement the above described steps with realizations x_i of $\boldsymbol{X}_i$ (in place of the observation x), we would arrive at the r-th step at a c.p.d. $P^{\boldsymbol{\theta}_r \,|\, \boldsymbol{X}_r = x_r}$. If it is Borel measurable as a function of $x^{(r)} = (x_1, \ldots, x_r)$ it would be a determination of the c.p.d $P^{\boldsymbol{\theta}^{(r)} \,|\, \boldsymbol{X}^{(r)} = x^{(r)}}$ (see Lemma A.1 in the appendix). Therefore an alternative noniterative description of our approach would be to take as the r-th posterior a determination of this c.p.d. at the special value $x^r := (x, \ldots, x)$, that is, setting the values $x_1, \ldots, x_r$ identical to our observation x:

$$\xi_x^{(r)} = P^{\boldsymbol{\theta}^{(r)} \,|\, \boldsymbol{X}^{(r)} = x^r}. \tag{2.5}$$

In the following we will choose this description of the r-th posterior, which again will depend on the chosen determination of the c.p.d., moreover in view of the fact, that with exception of some special cases, the set $A^{(r)} = \{x^r \,|\, x \in \mathcal{X}\}$ of all possible values x^r will have zero marginal probability $P^{\boldsymbol{X}^{(r)}}(A^{(r)}) = 0$.

An appealing possibility to reach a uniqueness of the r-th posterior for all $x \in \mathcal{X}$ appears, when the restriction to a continuous function of $x^{(r)}$ leads to a unique determination of the c.p.d. $P^{\boldsymbol{\theta}^{(r)} \,|\, \boldsymbol{X}^{(r)} = x^{(r)}}$. This determination would be especially interesting and therefore also its uniquely defined value $\xi_x^{(r)}$ at $x^{(r)} = x^r$.

Assuming $\mathcal{X} = \mathbb{R}^n$ a sufficient condition leading to the uniqueness of a continuous determination of the c.p.d. $P^{\boldsymbol{\theta}^{(r)} \,|\, \boldsymbol{X}^{(r)} = x^{(r)}}$ is the following (see Lemma A.2 in the appendix):
The marginal p.d.

$$P^{\boldsymbol{X}^{(r)}} = \sum_{m \in M} \pi_m Q_m \qquad (|M| < \infty) \tag{2.6}$$

of $\boldsymbol{X}^{(r)}$ is a finite mixture of p.d.'s Q_m, each having positive continuous density w.r.t. the Lebesgue measure λ_m over some linear space $L_m \subset \mathbb{R}^{nr}$. Moreover the p.d.'s Q_m are concentrated on disjoint sets $\mathcal{X}_m \subset L_m \quad : \quad Q_m(\mathcal{X}_m) = 1$.
An analogous remark applies, if there is interest in a conditional mean

$$E(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X}^{(r)} = x^r) =: \hat{\theta}_r(x) \tag{2.7}$$

as Bayes estimate of θ w.r.t. the posterior $\xi_x^{(r)}$. Under (2.6) this estimate would be uniquely defined for all $x \in \mathbb{R}^n$ under the restriction, that a continuous determination of the conditional mean $E(\boldsymbol{\theta}^{(r)} \mid \boldsymbol{X}^{(r)} = x^{(r)})$ is chosen.

Now we may consider the behavior of estimates $\hat{\theta}_r(x)$ for the interesting limit $r \to \infty$, even in the case of nonconvergence or of fluctuating sequences $\hat{\theta}_r(x)$.

Definition 2.1

If a limit $\lim_{r\to\infty} \hat{\theta}_r(x) = \theta(x)$ exists, then it is called a selfinformative (Bayes) limit. A set $\Theta(x) \subset \bar{\Theta}$ is called a weak selfinformative posterior carrier, if for all neighborhoods U of $\Theta(x)$ (that is open sets U containing $\Theta(x)$), it holds

$$\lim_{r\to\infty} \xi_x^{(r)}(U) = 1. \tag{2.8}$$

$\Theta(x)$ is a q-th order selfinformative posterior carrier, if

$$\lim_{r\to\infty} \int d[\theta, \Theta(x)]^q \, \xi_x^{(r)}(d\theta) = 0 \tag{2.9}$$

using the metric d in Θ and

$$d[\theta, T] = \inf_{\tau\in T} d(\theta, \tau). \tag{2.10}$$

$\Theta(x)$ is a selfinformative (Bayes) limit set, if

$$\lim_{r\to\infty} d[\hat{\theta}_r(x), \Theta(x)] = 0. \tag{2.11}$$

(If $\Theta(x) = \{\theta(x)\}$ in (2.11), then $\theta(x)$ is a selfinformative limit.)

Assume the special case of model (2.1) with realizations in $\mathfrak{X} = R^n$, p.d.'s P_θ dominated by a σ-finite measure and a finite dimensional parameter θ. The results in Bunke, Hennig, and Schmidt (1976) show, that under some weak regularity conditions ensuring the uniqueness of the MLE, the selfinformative limit exists and is identical to the MLE. This underlines the intuitive justification of selfinformative limits as convenient estimators in noninformative situations. It arises the interesting question, if such an equivalence extends to more general cases in which the MLE is not unique or when the MLE is not defined, as it may be the case in non- and semiparametric models. We will treat the dominated case with a possibly nonuniquely determined MLE in the following section 3.

3 Maximum likelihood and selfinformative limits

We assume a model (2.1) with p.d.'s P_θ having densities p_θ w.r.t. the σ-finite measure μ. As a determination of the r-th posterior w.r.t. the prior ξ we take the p.d. $\xi_x^{(r)}$ defined for $B \in \mathcal{B}$ by

$$\xi_x^{(r)}(B) = \int_B [p_\theta(x)]^r \xi(d\theta) \cdot \Big[\int_\Theta [p_\tau(x)]^r \xi(d\tau)\Big]^{-1} \quad (r > r_0) \tag{3.1}$$

under

Assumption A: For all $r \geq r_0$ ($r_0 \geq 0$) the integrals in (3.1) are finite.

Let $x \in \mathcal{X}$ be a fixed observation for which a MLE exists. The set of all MLE is

$$\hat{\Theta}(x) = \{\theta \in \Theta \,|\, p_\theta(x) = \max_{\tau \in \Theta} p_\tau(x)\}. \tag{3.2}$$

We will investigate, when a set $\Theta(x)$ is a selfinformative posterior carrier or even a limit set in the sense of Definition 2.1. The following assumptions will be sufficient:

Assumption B: For each neighborhood $U \neq \Theta$ of the closed set $\Theta(x)$ there is an neighborhood V of $\Theta(x)$ with $V \subset U$ such that $\xi(V) > 0$ and

$$\sup_{\theta \in U^c} p_\theta(x) < \inf_{\theta \in V} p_\theta(x) \qquad (U^c := \Theta \backslash U). \tag{3.3}$$

Remark 3.1

The assumption B is obviously fulfilled for the set $\hat{\Theta}(x)$ of MLE's, if $\xi(V) > 0$ for all nonempty open sets V, Θ is compact and $p_\theta(x)$ is continuous on Θ for fixed x. It is also fulfilled, if the assumption of compactness is replaced by local compactness and the property

$$\lim_{\theta \to \infty} p_\theta(x) = 0. \tag{3.4}$$

Then we may extend $p_\theta(x)$ to a continuous function on the compactified space Θ^∞. Replacing Θ by Θ^∞ in (3.2) does not change the set $\hat{\Theta}(x)$ of MLE's, which all remain in Θ, assuming w.l.o.g. that $p_\theta(x) > 0$ for some θ. The set $\hat{\Theta}(x)$ will obviously be compact.

Assumption C: Θ is a Banach space and $\int \|\theta\|^q \xi(d\theta) < \infty$ for a $q \geq 1$.

Theorem 3.1
(1) A set $\Theta(x)$ is under assumptions A, B a weak selfinformative posterior carrier. (2) Under the assumptions A,B,C a bounded set $\Theta(x)$ is a q-th order selfinformative posterior carrier. If $\Theta(x) = \{\theta(x)\}$ is a singleton, then $\theta(x)$ is a selfinformative limit.

(proofs of theorems are given in the appendix).

Because of this theorem and remark 3.1. many usual models for discrete or absolutely continuous variables X with a probability function or density being continuous in a k-dimensional parameter fulfil the assumptions of the theorem. If the prior has a finite first order moment, then the set of MLE's is a first order selfinformative carrier. If the MLE is unique, it is the selfinformative limit $\lim_{r\to\infty} \hat{\theta}_r(x) = \hat{\theta}(x)$, which obviously is identical for all such priors.

Remark 3.2. A semiparametric mixture model

A further interesting case, in which the set $\hat{\Theta}(x)$ of MLE's is at least a selfinformative posterior carrier, is a semiparametric mixture model (2.1). Here the density of P_θ w.r.t. a measure μ is given by

$$p_\theta(x) = \int_H f(x \,|\, \tau, \eta) \quad G(d\eta) \,, \tag{3.5}$$

with the unknown parameter $\theta = (\tau, G) \in \Xi \times \mathcal{G}$.

The function $f(\cdot \,|\, \tau, \eta)$ in (3.5) is assumed to be a density for fixed $\tau \in \Xi, \eta \in H$. Ξ and H are locally compact metric spaces, $f(x \,|\, \tau, \eta)$ is continuous on $\Xi \times H$ for fixed x and fulfills

$$\lim_{\underline{\tau}\to\infty} f(x|\underline{\tau}, \eta) \;=\; \lim_{\underline{\eta}\to\infty} f(x \,|\, \tau, \underline{\eta}) = 0 \quad , \quad (\tau \in \Xi, \eta \in H). \tag{3.6}$$

$\mathcal{G}$ is the set of all p.d.'s on the σ-Algebra of Borel sets of H.

It may be easily seen, that the density (3.5) fulfils the assumptions of remark 3.1, because $\mathcal{G}$ may be interpreted as a compact subset of the Banach space $\mathcal{L}$ of linear functionals L on the set C of bounded continuous functions on H:

$$Gc = \int c(\eta) G(d\eta) \qquad c \in C, G \in \mathcal{G} \tag{3.7}$$

The norms in L and C are:

$$\|L\| = \sup_{c \in C : \|c\| = 1} Lc \quad , \quad \|c\| = \sup_{\eta \in H} |\, c(\eta)\, |. \tag{3.8}$$

Now if Ξ is a Banach space with norm $\|\cdot\|$ we may see Θ as a locally compact subset of the Banach space $\Xi \times \mathfrak{L}$ with norm $\|(\tau, L)\| = \|\tau\| + \|L\|$. The assumption B is fulfilled with $q = 1$ for all priors ξ with a finite marginal 1. order moment of τ:

$$\int \|\theta\| \xi(d\theta) = \int \|\tau\| \xi(d\theta) + 1 < \infty \tag{3.9}$$

Therefore the set $\hat{\Theta}(x)$ of MLE's will be a selfinformative limit set under the above assumption.

The result in theorem 3.1 on a weak selfinformative carrier may be extended to more general cases. We assume a model (2.1) and a determination ξ_x of the c.p.d. $P^{\boldsymbol{\theta} \,|\, \boldsymbol{X}=x}$ which is weakly continuous in x and means, that $\int f d\xi_x$ is a continuous function in x for each bounded continuous function f.

Let x be an observation with $P^{\boldsymbol{X}}(S_x^\varepsilon) > 0$ for all spheres $S_x^\varepsilon = \{y \in \mathcal{X} \,|\, d(x,y) < \varepsilon\}$ and $\varepsilon > 0$. We remark, that then for all r the rectangle $R_{x^r}^\varepsilon := S_x^\varepsilon \times \cdots \times S_x^\varepsilon \in \mathcal{X}^r$ has a positive $\boldsymbol{X}^{(r)}$-marginal probability and at least $P^{\boldsymbol{X}}$-almost all x fulfil this assumption. For each $B \in \mathfrak{B}$ and $\varepsilon > 0$ we define a measure on Θ with

$$P_{x\varepsilon}^r(B) := \frac{P^{\boldsymbol{\theta}, \boldsymbol{X}^{(r)}}(B \times R_{x^r}^\varepsilon)}{P^{\boldsymbol{X}^{(r)}}(R_{x^r}^\varepsilon)}.$$

Assumption D: For all r exists a determination of the c.p.d. $P^{\boldsymbol{\theta} \,|\, \boldsymbol{X}^{(r)} = x^{(r)}}$ which is weakly continuous in $x^{(r)}$.

Under assumption D for $\varepsilon \downarrow 0$ the sequence of probability measures $P_{x\varepsilon}^r$ converge weakly for all r to the uniquely defined value $\xi_x^{(r)}$ at $x^{(r)} = x^r$ of the continuous determination of the c.p.d. $P^{\boldsymbol{\theta} \,|\, \boldsymbol{X}^{(r)} = x^{(r)}}$ (see Lemma A.3 in the appendix). Furthermore we will need for all $x \in \mathcal{X}$ following limit

$$p_{\theta,\tilde{\theta}}(x) = \limsup_{\varepsilon \downarrow 0} \frac{P_\theta(S_x^\varepsilon)}{P_{\tilde{\theta}}(S_x^\varepsilon)}. \tag{3.10}$$

If $P_{\tilde{\theta}}(S_x^\varepsilon) = 0$ we set the ratio to $= +\infty$, or $= 0$, if $P_\theta(S_x^\varepsilon) > 0$, or $= 0$, respectively. The limits $p_{\theta,\tilde{\theta}}$ are defined and measurable (see Hahn and Rosenthal (1948)) and

given by

$$p_{\theta,\tilde{\theta}}(x) = \frac{f_{\theta,\tilde{\theta}}(x)}{1 - f_{\theta,\tilde{\theta}}(x)}$$

with a determination $f_{\theta,\tilde{\theta}}(x)$ of the Radon Nikodym density of P_θ w.r.t. to $P_\theta + P_{\tilde{\theta}}$.

Assumption E: We assume $\Theta(x) \neq \Theta$. For each neighborhood U of $\Theta(x)$ and for each neighborhood $U \neq \Theta$ of $\Theta(x)$ there is an neighborhood V of $\Theta(x)$ with $V \subset U$ and $\xi(V) > 0$ such that

1. there is an $\varepsilon_0 > 0$ such for all $\varepsilon \in (0, \varepsilon_0)$ and for all $\theta \in U^c$ and $\tilde{\theta} \in V$ holds

$$P_\theta(S_x^\varepsilon) \leq P_{\tilde{\theta}}(S_x^\varepsilon), \tag{3.11}$$

2. the functions $p_{\theta,\tilde{\theta}}$ fulfil

$$\sup_{\theta \in U^c} \sup_{\tilde{\theta} \in V} p_{\theta,\tilde{\theta}}(x) < 1. \tag{3.12}$$

Theorem 3.2
Under assumption D and E a closed set $\Theta(x)$ is a weak selfinformative carrier.

4 Normal linear model with possibly unidentifiable parameters

Assume a normal linear model for the n dimensional observation y

$$y \sim N_n(X\beta, \sigma^2\Lambda), \tag{4.1}$$

where X is a fixed $n \times k$ matrix of rank d ($d \leq k \leq n$) and $\beta \in \mathbb{R}^k$, $\sigma > 0$ are unknown parameters. A version of the in general nonunique MLE of β and of σ^2 is

$$\hat{\beta} = (X^t\Lambda^{-1}X)^+X^t\Lambda^{-1}y \tag{4.2}$$

$$\hat{\sigma}^2 = n^{-1}(y - X\hat{\beta})^t\Lambda^{-1}(y - X\hat{\beta}). \tag{4.3}$$

In a Bayes approach the possible nonidentifiatibility of β plays no rule under a prior ξ for (β, σ^2). It is even possible to cover a further unknown parameter $\alpha \in \mathbb{R}^q$ from which the p.d. of y does not depend. This could e.g. be interesting in regression models with regression coefficients depending on time in a certain time interval and with observations at some subset of this time interval. The parameter α would contain the values of the regression coefficients for the times without observations. We assume a normal-gamma prior p.d. ξ defined by the following assumptions:

1. The marginal p.d. of σ^{-2} is a Gamma p.d. $\Gamma(\frac{w-k}{2}, s_\xi^{-1})$ with $w > k + 2$ and $s_\xi > 0$.

2. Under the condition of a fixed σ^2 the c.p.d. of $\gamma = (\alpha^t, \beta^t)^t$ is the normal p.d. $N_{q+k}(c_\xi, \sigma^2\Gamma_\xi^{-1})$ with $c_\xi = (a_\xi^t, b_\xi^t)^t \in \mathbb{R}^{q+k}$ and a positive definite Γ_ξ.

With $Z = (0 \vdots X)$ we have $y \sim N_n(Z\gamma, \sigma^2\Lambda)$. The posterior p.d. is then known to be again a normal-gamma p.d. and the posterior means are (see Humak (1977)):

$$\tilde{\gamma} = (\Gamma_\xi + G)^{-1}(\Gamma_\xi c_\xi + G\hat{\gamma}), \tag{4.4}$$

where $G = Z^t\Lambda^{-1}Z$,

$$\hat{\gamma} = (Z^t\Lambda^{-1}Z)^+ Z^t\Lambda^{-1}y = \begin{pmatrix} 0 \\ \hat{\beta} \end{pmatrix} \tag{4.5}$$

and

$$\tilde{\sigma}^2 = (w + n - k - q - 2)^{-1}[s_\xi + n\hat{\sigma}^2 + c_\xi^t\Gamma_\xi c_\xi + \hat{\gamma}^t G\hat{\gamma} - \tilde{\gamma}^t(\Gamma_\xi + G)\tilde{\gamma}]. \tag{4.6}$$

To calculate selfinformative limits we take r independent replications of observations $y^{(1)}, \ldots, y^{(r)}$ following the model (4.1), determine the corresponding posterior means (4.4), (4.6) and take $y^{(1)} = \cdots = y^{(r)} = y$. We obtain the posterior means

$$\tilde{\gamma}_r = (\Gamma_\xi + rG)^{-1}(\Gamma_\xi c_\xi + rG\hat{\gamma}), \tag{4.7}$$

$$\tilde{\sigma}_r^2 = (w + rn - k - q - 2)^{-1}[s_\xi + rn\hat{\sigma}^2 + c_\xi^t\Gamma_\xi c_\xi + r\hat{\gamma}^t G\hat{\gamma} - \tilde{\gamma}_r^t(\Gamma_\xi + rG)\tilde{\gamma}_r]. \tag{4.8}$$

Taking the limit $r \to \infty$ gives (see Lemma A.4 in the appendix):

$$\lim_{r\to\infty} \tilde{\gamma}_r = \hat{\gamma} + g_\xi, \quad \lim_{r\to\infty} \tilde{\sigma}_r^2 = \hat{\sigma}^2, \tag{4.9}$$

where

$$g_\xi = (I - G^+G)c_\xi = \begin{pmatrix} a_\xi \\ Qb_\xi \end{pmatrix}, \tag{4.10}$$

$$Q = (I - (X^t\Lambda^{-1}X)^+ X^t\Lambda^{-1}X). \tag{4.11}$$

Together with $\hat{\sigma}^2$ the limit $\hat{\gamma} + g_\xi$ is also a MLE (and LSE), because g_ξ is an element of the null space of G. We see that the selfinformative limit $\hat{\beta} + Qb_\xi$ for β will in general depend on the prior, but it is independent of ξ in the full rank case $d = k$, where $Q = 0$.

A further insight into the behaviour for $r \to \infty$ is given by the posterior marginal p.d. of γ. It is (see Humak (1977)) a generalized Student p.d. with $w + rn$ degrees of freedom, mean $\hat{\gamma} + g_\xi$ and covariance matrix of the form

$$C_r = D(\gamma|y) = \tilde{\sigma}_r^2(\Gamma_\xi + rG)^{-1}. \tag{4.12}$$

The reasoning in the proof of Lemma A.4 and (4.9) shows, that with $H = \Gamma_\xi^{-\frac{1}{2}}$

$$(4.13)\quad C_\xi = \lim_{r\to\infty} C_r = \hat{\sigma}^2 H[I-(HGH)^+(HGH)]H = \hat{\sigma}^2(I-G^+G)\Gamma_\xi^{-1}(I-G^+G).$$

Therefore the posterior marginal p.d. converges weakly to a normal p.d. with mean $\hat{\gamma}+g_\xi$ and covariance matrix C_ξ. This matrix generates the linear space L which is the null space of G. Obviously the set of all MLE's of γ under the model $y \sim N_n(Z\gamma, \sigma^2\Lambda)$ is the affine space

$$(4.14)\qquad A = \{\hat{\gamma}+l \,|\, l \in L\} = \{\begin{pmatrix} \alpha \\ \hat{\beta}+\beta \end{pmatrix} \,|\, \alpha \in R^q \quad , \quad X\beta = 0\}.$$

It is the carrier of the p.d. $N(\hat{\gamma}+g_\xi, C_\xi)$ and a weak selfinformative carrier because of the mentioned weak convergence of the posterior p.d.'s.

5 Estimation of the hazard rate

1. Discrete time and right censoring

We assume i.i.d. random variables $X_1, \ldots, X_n$ with values in $\mathfrak{X} = \{0, b, 2b, \ldots\}$ $(b > 0)$ and probabilities $f(m) = P(X_i = mb)$. The hazard rate h is then defined by

$$(5.1)\qquad h_f(m) = f(m) \quad / \quad \sum_{j=m}^{\infty} f(j).$$

The censoring is determined by a sequence $c_i \in (0, \infty)$ and leads to the observations

$$(5.2)\qquad T_i = \min\{X_i, c_i\} \quad , \quad \delta_i = I_{[0,c_i]}(X_i),$$

The vector Y of observations $Y_i = (X_i, \delta_i)$ has a probability function $P_f(y)$, which depends on the probability function f. The MLE $\hat{f}$ for f leads to

$$(5.3)\qquad \hat{h}(m) = h_{\hat{f}}(m) = \Delta N(m) \quad / \quad R(m),$$

with the "counting process"

$$(5.4)\qquad N(m) = \sum_{i=1}^{n} \delta_i \; I_{[0,mb]}(T_i),$$

(5.5) $$\Delta N(m) = N(m) - N(m-1)$$

and the "risk process"

(5.6) $$R(m) = \sum_{i=1}^{n} I_{[mb,\infty)}(T_i),$$

see Hjort (1990). Hjort also presents there the posterior means

(5.7) $$\tilde{h}(m) = E(h(m) \,|\, Y = y) = \frac{\Delta(m) + c(m)h_0(m)}{R(m) + c(m)}$$

in a Bayes approach as an alternative estimate. He assumes a prior, under which the values $h(m) \quad (m = 1, 2, \ldots)$ are i.i.d. with a Beta distribution

(5.8) $$h(m) \sim \text{Beta}\Big[c(m)h_0(m), c(m)(1 - h_0(m))\Big].$$

This prior determines a prior ξ over the class $\mathcal{F}$ of probability functions f.

To determine a selfinformative limit we consider X_i and constants $c_i \quad (i = 1, \ldots, rn)$ leading to observations (5.2). The r-th iterated Bayes estimate $\hat{h}_r$ will be given by

(5.9) $$\tilde{h}_r(m) = \frac{\Delta N_r(m) + c(m)h_0(m)}{R_r(m) + c(m)},$$

where N_r and R_r are given by (5.4) and (5.6) resp. substituting n by rn.
Assuming now, that the values $X_i, c_i \ (i = 1, \ldots, n)$ appear r-times we are lead to the r-th iterated Bayes estimates

(5.10) $$\hat{\theta}_r(m) = \frac{r\Delta N_r(m) + c(m)h_0(m)}{rR_r + c(m)}.$$

Their limit for $r \to \infty$ is just the MLE (5.3).

2. Continuous time and right censoring

We again assume censored observations (5.2) i.i.d. random variables $X_1, \ldots X_n$ having values in $\mathcal{X} = [0, \infty)$ and a p.d. P. The cumulative hazard function is defined by

$$H(t) = \int_0^t \frac{1}{P([s,\infty))} P(d\,s). \tag{5.11}$$

The GMLE in the sense of Gill (1989) is

$$\hat{H}(t) = \int_0^t \frac{1}{R(s)} N(d\,s), \tag{5.12}$$

where

$$N(s) = \sum_{i=1}^{n} \delta_i I_{[0,s]}(T_i) \tag{5.13}$$

$$R(s) = \sum_{i=1}^{n} I_{[s,\infty)}(T_i) \tag{5.14}$$

Hjort has derived posterior means $\tilde{H}$ under a Beta process with parameters c and h_0 (see Hjort (1990) for details) as a prior for H:

$$\tilde{H}(t) = \int_0^t \frac{c(s)}{C(s)+R(s)} H_0(d\,s) + \int_0^t \frac{1}{C(s)+R(s)} N(d\,s). \tag{5.15}$$

Again taking r replications of the model and assuming that the n same observations X_i and censoring times c_i occur r times leads to the r-th iterated Bayes estimate

$$\tilde{H}_r(t) = \int_0^t \frac{c(s)}{c(s)+rR(s)} H_0(d\,s) + \int_0^t \frac{1}{c(s)+rR(s)} rN(d\,s). \tag{5.16}$$

The selfinformative limit $r \to \infty$ of $\tilde{H}_r$ is again the GMLE (5.12).

6 Semiparametric multivariate linear models

In this section we present selfinformative limits in the case of the semiparametric linear model investigated by Bunke (2002) for n independent p-dimensional observations X_i with means and covariance matrices

(6.1) $$E(X_i) = z_i B \qquad D(X_i) = \Sigma \quad (i = 1, \ldots, n),$$

where z_i are the rows of a known $n \times k$ matrix Z of rank k. The $k \times p$ matrix B and Σ are unknown parameters. The model (6.1) is semiparametric, when even for known B, Σ the distributions P_i of X_i are unknown. To define the prior ξ for a Bayes approach we write

(6.2) $$X_i = z_i A + U_i \Lambda^{-1/2}$$

and assume that

(i) the vector $\mathbb{1} = (1, \ldots, 1)^t$ is contained in the linear space $R(Z)$ generated by the columns of Z.

(ii) the p-dimensional row vectors $U_1, \ldots, U_n$ are i.i.d. with p.d. G, under the condition of fixed $\theta = (A, \Lambda)$ and G.

(iii) under the prior ξ the parameter θ and G are random variables, which are independent,

(iv) the random p.d. G has a Dirichlet p.d. D_α where α is a measure on $\mathfrak{L}^p$ leading to the standard normal p.d. as the mean p.d. $E(G)$:

(6.3) $$a = \alpha(\mathbb{R}^p) \quad , \ \beta = a^{-1}\alpha = N(0,1) = E(G)$$

(v) the random parameter θ has a Normal-Wishart p.d. with density

(6.4) $$f(\theta) \propto |\Lambda|^{\tau/2}\, e[\Lambda(S_0 + (A - A_0)^t \Gamma_0 (A - A_0)],$$

where $\tau > k, \quad e[M] = \exp[-\mathrm{tr}M/2]$,
A_0 is a $k \times p$ matrix and Γ_0, S_0 are positive definite symmetric matrices.

We remark, that the distributions $P_i = P_i(A, \Lambda, G)$ of the observations given by (6.2) depend on θ and G.

Theorem 2.4 in Bunke (2002) yields expressions for the posterior means under the observation matrix X with rows X_i:

(6.5) $$\tilde{B} = E(B|X) \quad , \quad \tilde{\Sigma} = E(\Sigma|X) \quad , \quad \tilde{P}_i = E(P_i|X).$$

These expressions are extremely complicated and therefore we will present here only their general structure.
Let D be the set of all partitions v of the index set $\{1,\dots,n\}$. The posterior means are mixtures

$$\tilde{B} = \sum_{v\in D} k_v \tilde{B}_v \quad , \quad \tilde{\Sigma} = \sum_{v\in D} k_v \tilde{\Sigma}_v \quad , \quad \tilde{P}_i = \sum_{v\in D} k_v \tilde{P}_{iv} \tag{6.6}$$

with certain nonnegative weights k_v and $\sum\limits_{v\in D} k_v = 1$.
To obtain selfinformative limits we take the case of r independent replications of the linear model (6.1), that is we have a $rn \times p$ observation matrix $X_{(r)}$ with

$$E(X_{(r)}) = Z_{(r)}B \quad , \quad D(X_{(r)}) = D(\mathrm{vec}X_{(r)}) = \Sigma \otimes I_{rn}, \tag{6.7}$$

where

$$X^t_{(r)} = (X^t_{1,r} \vdots \dots \vdots X^t_{r,r}) \quad , \quad Z_{(r)} = \mathbb{1}_r \otimes Z \tag{6.8}$$

and where $\mathrm{vec}A = (a^t_1,\dots,a^t_n)^t$ denotes the vector of columns a_i of the matrix A and $\otimes$ is the Kronecker product. The detailed expressions for the conditional means (6.5) obtained for the observation matrix $X_{(r)}$ show, that they are continuous in $X_{(r)}$. In Bunke (2002) it is proved, that the marginal distribution p.d. P^X in the Bayes approach with observation matrix X and the prior ξ is a mixture

$$P^X = \sum_{v\in D} k_v P^X_v. \tag{6.9}$$

The p.d.'s P^X_v be concentrated at sets $\mathfrak{X}_v$ ($P^X_v(\mathfrak{X}_v) = 1$), which are disjoint and contained in linear spaces $L_{(v)} \subset \mathbb{R}^{n\times p}$. Each P^X_v is absolutely continuous w.r.t. the Lebesgue measure λ_v on $L_{(v)}$, so that the expressions (6.6) give in the case of an observation matrix $X_{(r)}$ the uniquely determined conditional means which are continuous in $X_{(r)}$ (see Lemma A.2).

Taking the original observation X obeying the model (6.1) for a special value $X_{(r)} = \mathbb{1}_r \otimes X$ in the model with r replicated observations leads to the r-th iterated Bayes estimates $\tilde{B}_{(r)}, \tilde{\Sigma}_{(r)}, \tilde{P}_{i(r)}$. The following theorem is proven in Johannes (2002) and shows, that their limits for $r \to \infty$ exist. These selfinformative limits in the sense of Definition 2.1. have an appealing "standard" form.

Theorem 6.1
The selfinformative limits of the iterated Bayes estimates are

$$\overline{B} = (Z^t Z)^{-1} Z^t X \tag{6.10}$$

$$\overline{\Sigma} = n^{-1} X^t [I_n - Z(Z^t Z)^{-1} Z^t] X \tag{6.11}$$

(6.12) $\overline{P}_i$: *empirical p.d. of the pseudoobservations* $X(i,j)$ $(j = 1, \ldots, n)$,

where

$$X(i,j) := X_j + (z_i - z_j)\overline{B}. \tag{6.13}$$

The estimates (6.10) and (6.11) are just the MLE under a parametric model (6.1) with a normal p.d.

But a more interesting property would be to be a generalized MLE in our semiparametric model (6.1) (or equivalently (6.2)) with unknown p.d. G. As we will see in the next section, this is true only in special cases.

7 Generalized maximum likelihood estimates

In the following we still assume the general model (2.1) and introduce the definition of Kiefer and Wolfowitz (1956) for a generalized MLE. We use for all $\theta, \tilde{\theta} \in \Theta$ a determination $f_{\theta,\tilde{\theta}}$ of the density of P_θ w.r.t. $P_\theta + P_{\tilde{\theta}}$.

Definition 5.1
For the observation $x \in \mathcal{X}$ *the value* $\hat{\theta} = \hat{\theta}(x) \in \Theta$ *is called a generalized maximum likelihood estimate (GMLE) of* θ*, if*

$$f_{\hat{\theta},\theta}(x) \geq f_{\theta,\hat{\theta}}(x) \quad \textit{for all} \quad \theta \in \Theta. \tag{7.1}$$

The GMLE depend on the chosen density determinations $f_{\theta,\tilde{\theta}}$. The ordinary MLE under a dominated class of p.d.'s P_θ is a special case. There are other alternative approaches to the definition of a GMLE, e.g. see Gill (1989) or Grenander (1981).

The definition of Kiefer and Wolfowitz is especially appealing because of its simplicity and because appears to be the direct extension of the definition in the dominated case. Here we will not go into details of a comparison between different approaches to GMLE. We now take as a special case a semiparametric linear model

$$X_i = z_i B + U_i \qquad (i = 1, \ldots, n) \tag{7.2}$$

for p-dimensional random variables X_i with p.d. $P_{i\theta}$, where z_i are the rows of a known $n \times k$ matrix Z of rank $k < n$, B is an unknown $k \times p$ matrix and $U_1, \ldots, U_n$ are i.i.d. p-dimensional random variables with zero mean, finite second order moments and unknown p.d. G.

With

$$\mathcal{G} = \{G \mid \text{p.d. on } \mathfrak{L}^p \quad , \quad \int \|x\|^2 G(dx) < \infty\}, \tag{7.3}$$

$$\mathcal{G}_0 = \{G \in \mathcal{G} \mid \int x G(dx) = 0\} \tag{7.4}$$

we have a semiparametric model (2.1) for the observation matrix X with unknown parameter $\theta = (B, G) \in \Theta = \mathbb{R}^{k \times p} \times \mathcal{G}_0$.

The following theorem holds for a determination $f_{\theta,\tilde{\theta}}$ of the density of P_θ w.r.t. $P_\theta + P_{\tilde{\theta}}$ $(\theta \neq \tilde{\theta})$ obeying the natural equation

$$g_{\theta,\tilde{\theta}} = \prod_{i=1}^{n} \quad g_{i,\theta,\tilde{\theta}}(X_i) = f_{\theta,\tilde{\theta}}(X)\, h_{\theta,\tilde{\theta}}(X), \tag{7.5}$$

where $g_{i,\theta,\tilde{\theta}}$ is a determination of the density of $P_{i\theta}$ w.r.t. $P_{i\theta} + P_{i\tilde{\theta}}$ and $h_{\theta,\tilde{\theta}}$ is a determination of the density of $P_\theta + P_{\tilde{\theta}}$ w.r.t. $\bigtimes_{i=1}^{n}(P_{i\theta} + P_{i\tilde{\theta}})$. Obviously the equation (7.5) holds for $(P_\theta + P_{\tilde{\theta}})$-almost all X.

Theorem 7.1
A possibility nonunique GMLE $\hat{\theta}(X)$ in the model (7.2) is given by an element $\hat{B}(X)$ from

$$\{\hat{B} \in R^{k \times p} \mid \overline{X} = \overline{Z}\hat{B}, \quad g(\hat{B} \mid X) \geq g(B \mid X) \quad \forall\, B \in R^{k \times p}\} \tag{7.6}$$

and by

(7.7) $\hat{G}(x)$: *empirical p.d. of the residuals* $X_i - z_i\hat{B}(X) \quad (i = 1, \ldots, n)$,

where

$$g(B \,|\, X) = n^{-n} \prod_{i=1}^{n} \, |\, \{j \,|\, X_j - z_j B = X_i - z_i B\} \,|, \tag{7.8}$$

$$\overline{X} = n^{-1} \sum_{i=1}^{n} X_i \quad , \quad \overline{Z} = n^{-1} \sum_{i=1}^{n} z_i \tag{7.9}$$

Remark 5.1

The GMLE is unique for observations,

$$X \in R(Z) = \{ZB \,|\, B \in \mathbb{R}^{k \times p}\}. \tag{7.10}$$

because then $\hat{B}(X) = (Z^t Z)^{-1} Z^t X$ is the unique solution $\hat{B}$ of the equation $\overline{X} = \overline{Z}\hat{B}$ with

$$g(\hat{B} \,|\, X) = \max_B g(B \,|\, X). \tag{7.11}$$

In a special case of a location model

$$X_i = B + U_i \qquad U_i \sim G, \quad \text{i.i.d.}, \tag{7.12}$$

that is, with $k = 1$ and $Z = \mathbb{1}_n$, we obtain from theorem 7.1:

Theorem 7.2
The GMLE in the semiparametric location model (7.12) is unique and given by the sample mean

$$\hat{B} = \overline{X} = n^{-1} \sum_{i=1}^{n} X_i, \tag{7.13}$$

and by

(7.14) $\hat{G}$: *empirical p.d. of the residuals* $X_i - \overline{X} \quad (i = 1, \ldots, n)$.

We see, that the GMLE and the noninformative limits (see theorem 6.1) are identical. The following example shows, that this is not always the case in other linear models.

Example:
We consider a two-sample model with $p = 1$ and $n = 4$:

$$Z = \begin{pmatrix} \mathbb{1}_2 & 0 \\ 0 & \mathbb{1}_2 \end{pmatrix} \quad , \quad B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{7.15}$$

Take the observation $x = (1, -1, 2, 2)^t$. Then the selfinformative limit (6.10) will be $\overline{B} = (0, 2)^t$ and leads to the set $\{1; -1, 0, 0\}$ of residuals $X_i - z_i\overline{B}$. As it is easily seen the maximum of $g(B \,|\, X)$ is reached for $B^* = (-0.5, 2.5)$ with residual set $\{1.5, -0.5, -0.5, -0.5\}$ and also for $\tilde{B} = (0.5, 1.5)$ but not for $\overline{B}$:

$$g(\overline{B} \,|\, X) = 0.0156 < \max_B g(B \,|\, X) = g(B^* \,|\, X) = 0,1054. \tag{7.16}$$

Similar situations may be found for other observations $X \notin R(Z)$.

References

Bernardo, J. M., and A. F. Smith (1994): *Bayesian theory.* Wiley Series in Probability and Statistics. Chichester: Wiley. xiv, 586 p.

Billingsley, P. (1968): *Convergence of probability measures.* New York-London-Sydney-Toronto: John Wiley and Sons, Inc. XII, 253 p.

Box, G. E., and G. C. Tiao (1992): *Bayesian inference in statistical analysis.* Wiley Classics Library. New York, NY: Wiley. 606 p.

Bunke, H., C. Hennig, and W. Schmidt (1976): "Weighted combination of prior and sample information and parameter estimation.," *Math. Operationsforsch. Statistik*, 7, 665–672.

Bunke, O. (2002): "Bayes estimates in multivariate semiparametric linear models," Discussion paper 58, Sonderforschungsbereich 373, Humboldt University, Berlin.

Efron, B., and C. Morris (1973): "Combining possibly related estimation problems. Discussion.," *J. R. Stat. Soc., Ser. B*, 35, 379–421.

Gill, R. D. (1989): "Non- and semi-parametric maximum likelihood estimators and the von Mises method. I.," *Scand. J. Stat.*, 16(2), 97–128.

Grenander, U. (1981): *Abstract inference.* Wiley Series in Probability and Mathematical Statistics. New York etc.: John Wiley & Sons. IX, 526 p.

Hahn, H., and A. Rosenthal (1948): *Set functions.* Albuquerque, New Mexico: The University of New Mexico Press. IX, 324 p.

Hartigan, J. (1983): *Bayes theory.* Springer Series in Statistics. New York etc.: Springer-Verlag. XII, 145 p.

Hjort, N. L. (1990): "Nonparametric Bayes estimators based on beta processes in models for life history data.," *Ann. Stat.*, 18(3), 1259–1294.

Humak, K. (1977): *Statistische Methoden der Modellbildung Band I. Statistische Inferenz für lineare Parameter.* Mathematische Lehrbücher und Monographien. II. Abt. Band 43. Berlin: Akademie-Verlag. XVI, 516 S.

Johannes, J. (2002): "Verallgemeinerte Maximum-Likelihood-Methoden und der selbstinformative Grenzwert," Ph.D. thesis, Humboldt University, Berlin.

Kiefer, J., and J. Wolfowitz (1956): "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters.," *Ann. Math. Stat.*, 27, 887–906.

Lindley, D., and A. Smith (1972): "Bayes estimates for the linear model.," *J. R. Stat. Soc., Ser. B*, 34, 1–41.

A Appendix

Lemma A.1

Let $\boldsymbol{\theta}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2$ be random variables with values in Θ and $\mathfrak{X}$ resp. We assume :

1. $\boldsymbol{\theta} \sim \xi$, $\boldsymbol{\theta}_1 \sim \xi$
2. *the c.p.d.'s*

$$P^{\boldsymbol{Y}_1 \,|\, \boldsymbol{\theta}_1 = \theta} = P_\theta, \quad P^{\boldsymbol{X}_1, \boldsymbol{X}_2 \,|\, \boldsymbol{\theta} = \theta} = P_\theta \times P_\theta \tag{A.1}$$

3. *determinations of the c.p.d.'s*

$$\xi_x = P^{\boldsymbol{\theta}_1 \,|\, \boldsymbol{Y}_1 = x}, \quad \xi(x_1, x_2) = P^{\boldsymbol{\theta} \,|\, \boldsymbol{X}_1 = x_1, \boldsymbol{X}_2 = x_2} \tag{A.2}$$

4. *the c.p.d. $Q_x = P^{\boldsymbol{\theta}_2, \boldsymbol{Y}_2 \,|\, \boldsymbol{\theta}_1 = \theta, \boldsymbol{Y}_1 = x}$ is given for $B \in \mathfrak{B}$, $A \in \mathfrak{A}$ by*

$$Q_x(B \times A) = \int_B P_\theta(A) \xi_x(d\theta) \tag{A.3}$$

5. *$\nu(y_1, y_2)$ is a determination of the c.p.d $\xi(x_1, x_2) = P^{\boldsymbol{\theta}_2 \,|\, \boldsymbol{\theta}_1 = \theta, \boldsymbol{Y}_1 = y_1, \boldsymbol{Y}_2 = y_2}$*

Then it holds, that the marginal p.d.'s of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ are identical and

$$P(\xi(\boldsymbol{X}_1, \boldsymbol{X}_2)(B) = \nu(\boldsymbol{X}_1, \boldsymbol{X}_2)(B)) = 1 \qquad (B \in \mathfrak{B}). \tag{A.4}$$

Proof

Let $f : (\mathfrak{X} \times \mathfrak{X}, \mathfrak{A} \times \mathfrak{A}) \to (\mathbb{R}^1, \mathfrak{L}^1)$ be a bounded measurable function. Then the properties of c.p.d.'s give the mean

$$\begin{aligned} \mathbb{E}\, f(\boldsymbol{Y}_1, \boldsymbol{Y}_2) &= \int \int \Big[\int f P_{\theta_2}(dy_2) \Big] \xi_{y_1}(d\theta_2) P^{\boldsymbol{Y}_1}(dy_1) \\ &= \int \int \Big[\int f P_{\theta_2}(dy_2) \Big] P_{\theta_2}(dy_1) \xi(d\theta_2) = \mathbb{E}\, f(\boldsymbol{X}_1, \boldsymbol{X}_2) \end{aligned} \tag{A.5}$$

This proves, that the marginal p.d.'s of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ are identical. Using indicator functions I for sets $A_1, A_2 \in \mathfrak{A}$, $B \in \mathfrak{B}$ we obtain for the c.p.d. $\nu(x_1, x_2)(\cdot)$

from assumption 5. :

$$
\begin{aligned}
(A.6) \qquad & \int\int I_{A_1\times A_2}(x_1,x_2)\nu(x_1,x_2)(B)P^{\boldsymbol{X}_1,\boldsymbol{X}_2}(d(x_1,x_2)) = \\
& = \int\Big[\int I_{A_1\times A_2}(y_1,y_2)\nu(y_1,y_2)(B)P^{\boldsymbol{Y}_2\,|\,\boldsymbol{Y}_1=y_1}(dy_2)\Big]P^{\boldsymbol{Y}_1}(dy_1) = \\
& = \int\Big[\int\int I_{A_1\times A_2\times B}(y_1,y_2,\theta_2)P_{\theta_2}(dy_2)\xi_{y_1}(d\theta_2)\Big]P^{\boldsymbol{Y}_1}(dy_1) = \\
& = \int\int I_{A_1\times B}(y_1,\theta)P_\theta(A_2)\xi_{y_1}(d\theta)\Big]P^{\boldsymbol{Y}_1}(dy_1) = \\
& = \int\int I_{A_1\times B}(y_1,\theta)P_\theta(A_2)P_\theta(dy_1)\xi(d\theta) = \\
& = P^{\boldsymbol{\theta},\boldsymbol{X}_1,\boldsymbol{X}_2}(B\times A_1\times A_2).
\end{aligned}
$$

This shows, that ν is a determination of the c.p.d. $P^{\boldsymbol{\theta}\,|\,\boldsymbol{X}_1=x_1,\boldsymbol{X}_2=x_2}$ □

The next Lemma states sufficient conditions for the uniqueness of a continuous c.p.d. or conditional expectation.

Lemma A.2
Let $\boldsymbol{\theta},\boldsymbol{X}$ be random variables with values in Θ and $\mathcal{X}=\mathbb{R}^n$. We assume, that the marginal p.d. of $\boldsymbol{X}$ is a mixture

$$(A.7) \qquad Q=\sum_{m\in M}\pi_m Q_m \quad (\pi_m>0,\, m\in M)$$

of a finite number of p.d.'s Q_m concentrated at disjoint sets $\mathcal{X}_m\in\mathfrak{A}$:

$$(A.8) \qquad Q_m(\mathcal{X}_m)=1, \quad \mathcal{X}=\sum_{m\in M}\mathcal{X}_m, \quad \mathcal{X}_m\cap\mathcal{X}_{m'}=\emptyset \quad (m\neq m').$$

The sets $\mathcal{X}_m$ are contained in linear spaces $L_m\subset\mathbb{R}^n$ and moreover each p.d. Q_m has a positive continuous density q_m w.r.t. the Lebesgue measure λ_m on L_m.
Let $f:\Theta\to\mathbb{R}^1$ be a measurable function with $\mathbb{E}\,|f(\boldsymbol{\theta})|<\infty$. If there is a continuous determination of the conditional mean $g(x)=\mathbb{E}(f(\boldsymbol{\theta})|\,\boldsymbol{X}=x)$, then it is unique.

Proof
Let g,h be continuous determinations of $\mathbb{E}(f(\boldsymbol{\theta})|\,\boldsymbol{X}=x)$. Because of the definition of conditional means the measures μ and ν on $\mathfrak{A}=\mathfrak{L}^n$ given by

$$(A.9) \qquad \mu(A)=\int_A g(x)Q(dx), \quad \nu(A)=\int_A h(x)Q(dx),$$

are identical. If $x\in\mathcal{X}_m$ is fixed, their values are identical for the spheres

$$(A.10) \qquad S_m^\epsilon=\{y\in L_m\,|\,||y-x||<\epsilon\}.$$

Taking $l_m^\epsilon := [\lambda_m(S_m^\epsilon)]^{-1}$ it follows, that

$$\begin{aligned} \lim_{\epsilon\downarrow 0} l_m^\epsilon \mu(S_m^\epsilon) &= \pi_m \lim_{\epsilon\downarrow 0} l_m^\epsilon \int\limits_{S_m^\epsilon} g(x) q_m(x) \lambda_m(dx) = \pi_m g(x) q_m(x) \\ &= \lim_{\epsilon\downarrow 0} l_m^\epsilon \nu(S_m^\epsilon) = \pi_m h(x) q_m(x) \end{aligned} \tag{A.11}$$

and therefore $g(x) = h(x)$. □

Proof of theorem 3.1.

In the case $\Theta(x) = \Theta$ the set $\Theta(x)$ is obviously a weak and a q-th order selfinformative carrier. Assume now $\Theta(x) \neq \Theta$.

(1)

Let U, V be neighborhoods of $\Theta(x)$ with $V \subset U$, $U \neq \Theta$, $\xi(V) > 0$ and 3.3. Then we have for $r \to \infty$:

$$\xi_x^{(r)}(U^c) \leq \int\limits_{U^c} \Big[\int\limits_V \Big(\frac{p_\tau(x)}{p_\theta(x)} \Big)^r \xi(d\tau) \Big]^{-1} \xi(d\theta) \leq \Big[\frac{\inf\limits_{\tau\in V} p_\tau(x)}{\sup\limits_{\theta\in U^c} p_\theta(x)} \Big]^{-r} \frac{\xi(U^c)}{\xi(V)} \to 0. \tag{A.12}$$

(2)

Let θ_0 be a fixed element of $\Theta(x)$. Then there is a positive $C < \infty$ such that

$$d[\theta, \Theta(x)]^q \leq \Big(||\theta|| + ||\theta_0|| + \sup_{\tau\in\Theta(x)} ||\tau - \theta_0|| \Big)^q \leq C(||\theta||^q + 1). \tag{A.13}$$

For a fixed $\epsilon > 0$ there is a neighborhood U of $\Theta(x)$ with

$$\sup_{\theta\in U} d[\theta, \Theta(x)]^q \leq \frac{\epsilon}{3}. \tag{A.14}$$

Let V be a neighborhood of $\Theta(x)$ with $V \subset U$, $\xi(V) > 0$ and (3.3). Then with assumption C it follows for $r \to \infty$

$$\int\limits_{U^c} ||\theta||^q \xi_x^{(r)}(d\theta) \leq \Big[\frac{\inf\limits_{\tau\in V} p_\tau(x)}{\sup\limits_{\theta\in U^c} p_\theta(x)} \Big]^{-r} \frac{1}{\xi(V)} \int\limits_{U^c} ||\theta||^q \xi(d\theta) \to 0. \tag{A.15}$$

Therefore because of (A.12)-(A.15) there is a $r_\epsilon > 0$ such that for $r > r_\epsilon$

$$\int\limits_\Theta d[\theta, \Theta(x)]^q \xi_x^{(r)}(d\theta) \leq \frac{\epsilon}{3} + C \int\limits_{U^c} ||\theta||^q \xi_x^{(r)}(d\theta) + C\xi_x^{(r)}(U^c) < \epsilon. \tag{A.16}$$

This proves that $\Theta(x)$ is a q-th order selfinformative carrier.

Assume now $\Theta(x) = \{\theta(x)\}$. The preceding proof gives

$$\lim_{r\to\infty} \int ||\theta(x) - \theta|| \xi_x^{(r)}(d\theta) = 0 \tag{A.17}$$

and therefore $\int ||\theta|| \xi_x^{(r)}(d\theta) < \infty$ for sufficiently large r, so that then the mean $\hat{\theta}_r(x)$ is defined. Because of

$$||\hat{\theta}_r(x) - \theta(x)|| = || \int [\theta - \theta(x)] \xi_x^{(r)}(d\theta)|| \leq \int ||\theta - \theta(x)|| \xi_x^{(r)}(d\theta) \tag{A.18}$$

and (A.17) we obtain $\lim_{r\to\infty} \hat{\theta}_r(x) = \theta(x)$. □

Lemma A.3

Let $\boldsymbol{\theta}, \boldsymbol{X}$ be random variables with values in Θ and $\mathcal{X}$. We assume, that a determination $\xi_y^{(r)}$ of the c.p.d. $P^{\boldsymbol{\theta} \mid \boldsymbol{X}^{(r)}=y}$ exist, which is weakly continuous in y.
Then for all x with $P^{\boldsymbol{X}}(S_x^\varepsilon) > 0$ $(\epsilon > 0)$ and all $r \geq 1$ follows the weak convergence

$$P_{x\varepsilon}^r \xrightarrow{\mathcal{W}} \xi_{x^r}^{(r)}, \quad \textit{for } \epsilon \downarrow 0.$$

Proof

Let $f : \Theta \to \mathbb{R}^1$ be a bounded continuous function. Because of the definition of the conditional mean $g_f^r(y) = \int f d\xi_y^{(r)}$ it holds

$$\int_\Theta f(\theta) P_{x\varepsilon}^r(d\theta) = \frac{1}{P^{\boldsymbol{X}^{(r)}}(R_{x^r}^\varepsilon)} \int_{R_{x^r}^\varepsilon} g_f^r(y) P^{\boldsymbol{X}^{(r)}}(dy) \tag{A.19}$$

and obviously

$$\inf_{y \in R_{x^r}^\varepsilon} g_f^r(y) \leq \int f dP_{x\varepsilon}^r \leq \sup_{y \in R_{x^r}^\varepsilon} g_f^r(y). \tag{A.20}$$

Taking the limit $\varepsilon \downarrow$ in (A.20) proves the Lemma:

$$\lim_{\varepsilon \downarrow 0} \int f dP_{x\epsilon}^r = g_f^r(x^r). \tag{A.21}$$

□

Proof of theorem 3.2.

Let H be a neighborhood of $\Theta(x)$. Then there are disjoint neighborhoods Z of H^c and U of $\Theta(x)$. Let V be the neighborhood stated in assumption E. It follows for all r

$$\xi_{x^r}^r(H^c) \leq \xi_{x^r}^r(Z) \leq \liminf_{\epsilon \downarrow 0} P_{x\varepsilon}^r(Z) \leq \limsup_{\epsilon \downarrow 0} \int_Z \frac{[P_\theta(S_x^\epsilon)]^r}{\int_V [P_{\underline{\theta}}(S_x^\epsilon)]^r \xi(d\underline{\theta})} \xi(d\theta), \tag{A.22}$$

where the second inequality holds because of the weak convergence stated in Lemma A.3 (see Billingsley (1968)). Two applications of the inequality of Jensen yield

$$\xi_x^r(H^c) \leq \frac{1}{\xi(V)} \limsup_{\epsilon \downarrow 0} \int_Z \Big[\frac{1}{\xi(V)} \int_V \frac{P_\theta(S_x^\epsilon)}{P_{\underline{\theta}}(S_x^\epsilon)} \xi(d\underline{\theta}) \Big]^r \xi(d\theta). \tag{A.23}$$

With (3.11), we apply the theorem of dominated convergence together with (3.10) and with assumption (3.12) it follows

$$\lim_{r\to\infty} \xi_x^r(H^c) \leq \lim_{r\to\infty} \frac{\xi(Z)}{\xi(V)} \Big[\sup_{\theta\in Z} \sup_{\underline{\theta}\in V} p_{\theta,\underline{\theta}}(x) \Big]^r = 0. \tag{A.24}$$

□

Lemma A.4

Let Γ_ξ be a $n \times n$ positive definite matrix and G be a $n \times n$ positive semidefinite matrix with rank $q \leq n$, then it holds for every $a, c \in \mathbb{R}^n$, that

$$\lim_{r\to\infty} (r^{-1}\Gamma_\xi + G)^{-1}(r^{-1}\Gamma_\xi a + Gc) = (I_n - G^+G)a + G^+Gc \tag{A.25}$$

$$\lim_{r\to\infty} (r^{-1}\Gamma_\xi a + Gc)^t(r^{-1}\Gamma_\xi + G)^{-1}(r^{-1}\Gamma_\xi a + Gc) = c^tGc \tag{A.26}$$

Proof

We assume, that $\lambda_1, \ldots, \lambda_q$ are the positive eigenvalues of the positive semidefinite matrix $\Gamma_\xi^{-\frac{1}{2}} G \Gamma_\xi^{-\frac{1}{2}}$, the orthonormal vectors $u_1, \ldots, u_q$ are the eigenvectors and form together with the vectors $u_{q+1}, \ldots, u_n$ an orthonormal basis. Then it holds

$$(r^{-1}\Gamma_\xi + G)^{-1} = \Gamma_\xi^{-\frac{1}{2}} \Big(\sum_{i=q+1}^n r u_i u_i^t + \sum_{i=1}^q (r^{-1} + \lambda_i)^{-1} u_i u_i^t \Big) \Gamma_\xi^{-\frac{1}{2}}. \tag{A.27}$$

Therefore it follows

$$\begin{aligned}(r^{-1}\Gamma_\xi + G)^{-1}(r^{-1}\Gamma_\xi a + Gc) = \Gamma_\xi^{-\frac{1}{2}} \Big(\sum_{i=q+1}^n u_i u_i^t + \sum_{i=1}^q (r^{-1} + \lambda_i)^{-1} r^{-1} u_i u_i^t \Big) \Gamma_\xi^{-\frac{1}{2}} \Gamma_\xi a + \\ + \Gamma_\xi^{-\frac{1}{2}} \Big(\sum_{i=1}^q (r^{-1} + \lambda_i)^{-1} u_i u_i^t \Big) \Gamma_\xi^{-\frac{1}{2}} Gc \end{aligned} \tag{A.28}$$

and furthermore

$$\begin{aligned}\lim_{r\to\infty} (r^{-1}\Gamma_\xi + G)^{-1}(r^{-1}\Gamma_\xi a + Gc) &= \Gamma_\xi^{-\frac{1}{2}} \Big(\sum_{i=q+1}^n u_i u_i^t \Big) \Gamma_\xi^{-\frac{1}{2}} \Gamma_\xi a + \Gamma_\xi^{-\frac{1}{2}} \Big(\sum_{i=1}^q \lambda_i^{-1} u_i u_i^t \Big) \Gamma_\xi^{-\frac{1}{2}} Gc \\ &= (I_n - G^+G)a + G^+Gc. \end{aligned} \tag{A.29}$$

(A.26) is proved in an analogous manner. □

Proof of theorem 7.1.

Let X be a fixed observation. Because of (7.5) $\hat{\theta} = \hat{\theta}(X)$ is a GMLE if

$$g_{\hat{\theta},\theta}(X) \geq g_{\theta,\hat{\theta}}(X) \qquad \text{for all } \theta \in \Theta. \tag{A.30}$$

We remark, that we may extend the model (7.2) to p.d.'s G with possibly nonzero mean, replacing Θ by $\tilde{\Theta} = R^{n\times p} \times \mathcal{G}$ (see (7.3)). We extend also the density determinations g, f, h to determinations obeying (7.5) for $\theta, \tilde{\theta} \in \tilde{\Theta}$.

We now show, that for all $\theta = (A, G) \in \Theta$ there is a $\theta_A = (A, F_A) \in \tilde{\Theta}$ with

$$F_A = n^{-1} \sum_{i=1}^{n} \delta_{X_i - z_i A} \tag{A.31}$$

and

$$g_{\theta_A,\theta}(X) \geq g_{\theta,\theta_A}(X). \tag{A.32}$$

The density $g_{i,\theta_A,\theta}(Y)$ is obviously positive and uniquely determined for $Y = X_i$ and $\theta \in \tilde{\Theta}$ by

$$g_{i,\theta_A,\theta}(X_i) = \frac{P_{i\theta_A}(\{X_i\})}{P_{i\theta_A}(\{X_i\}) + P_{i\theta}(\{X_i\})}. \tag{A.33}$$

and therefore also

$$g_{i,\theta,\theta_A}(X_i) = \frac{P_{i\theta}(\{X_i\})}{P_{i\theta_A}(\{X_i\}) + P_{i\theta}(\{X_i\})}. \tag{A.34}$$

It is known, that the empirical p.d. F of observations $Y_i \quad (i = 1, \ldots, n)$ is a nonparametric GMLE in the sense

$$\prod_{i=1}^{n} F(\{Y_i\}) \geq \prod_{i=1}^{n} G(\{Y_i\}) \qquad \text{for all } G \in \mathcal{G}, \tag{A.35}$$

(see Kiefer and Wolfowitz (1956)).
Because of (7.8), (A.33) and (A.34) and

$$P_{i\theta_A}(\{X_i\}) = F_A(\{X_i - z_i A\}) \tag{A.36}$$

$$P_{i\theta}(\{X_i\}) = G(\{X_i - z_i A\}) \qquad (\theta = (A, G)) \tag{A.37}$$

$$g(B \mid X) = \prod_{i=1}^{n} F_B(\{X_i - z_i B\}) \qquad (B \in R^{k\times p}) \tag{A.38}$$

we have

$$q(\theta_B, \theta) = g_{\theta_B,\theta}(X)\Big[g_{\theta,\theta_B}(X)\Big]^{-1} = g(B \mid X)\Big[\prod_{i=1}^{n} G(\{X_i - z_i A\})\Big]^{-1}. \tag{A.39}$$

Now (A.39) and the inequality (A.35) have the consequence $q(\theta_A, \theta) \geq 1$ and as $g(A \mid X)$ remains unchanged replacing A by

$$\tilde{A} = A + (Z^t Z)^+ Z^t \mathbb{1}_n(\overline{X} - \overline{Z}A), \tag{A.40}$$

it follows that $g(\tilde{A} \mid X) = g(A \mid X)$ and

$$q(\theta_{\tilde{A}}, \theta) \geq 1 \qquad \text{for all} \quad \theta = (A, G) \in \Theta. \tag{A.41}$$

As $\overline{X} = \overline{Z}\tilde{A}$ holds, we have $g(\hat{B} \mid X) \geq g(\tilde{A} \mid X)$ for $\hat{\theta} = \hat{\theta}_{\hat{B}} = (\hat{B}, \hat{G})$ satisfying the conditions (e.g. (7.6)) of theorem 7.1. Therefore we have

$$q(\hat{\theta}, \theta) \geq q(\theta_{\tilde{A}}, \theta) \geq 1 \quad \text{for all} \quad \theta = (A, G) \in \Theta, \tag{A.42}$$

that is the inequality (A.30) is valid and $\hat{\theta}$ is a GMLE. □

Proof of theorem 7.2.

If we assume, that $Z = \mathbb{1}_n$, then it follows with theorem 7.1 and $\overline{z} = 1$, that

$$\hat{B} = \overline{X}, \quad \hat{G} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i - \overline{X}}. \tag{A.43}$$

Therefore the GMLE for the p.d. P of an observation X_i is the empirical p.d. $\hat{P}$ of the observations. Alternatively, if we have a nonparametric model, that means the observations $X_1, \ldots, X_n$ are independent and identically distributed and the unknown p.d. P is a member of the set $\mathcal{P}$ of all distributions, then the GMLE is the empirical p.d. $\hat{P}$ of the observations (see Kiefer and Wolfowitz (1956) and Gill (1989)). The semiparametric location model is the special case, where the p.d. P is a member of the set $\mathcal{P}_+$ of all distributions with finite first and second moment. The empirical p.d. is contained in the set $\mathcal{P}_+$ and therefore also the GMLE in the semiparameteric location model . □